Appendix D. Description of SIPP Panel File and Data Quality

DESCRIPTION OF SIPP 1984 PANEL FILE

The estimates presented in this report are based on the first SIPP panel file. This file contains monthly data for persons over a 32-month period. The staggered SIPP design (described in appendix A) means that the actual reference periods are June 1983 to January 1986, July 1983 to February 1986, August 1983 to March 1986, and September 1983 to April 1986.

Each person in the panel file has been assigned three weights: a weight for calendar year 1984, a weight for calendar year 1985 and a weight for the 32-month period. In order to receive a non-zero weight, a person must have an observation for each month of the relevant reference period (in this report, 1984 and 1985) or have a complete set of observations up until the time he or she died or became institutionalized. The data shown in this report would be affected if characteristics of persons with an incomplete set of observations differed from those with a complete set.

Table D-1 shows three categories of sample persons by sex, age, and program participation status. The numbers in the table are unit counts; they are not weighted. The category "complete set of interviews obtained" includes 32,391 persons, but 1, 307 of these persons died or were institutionalized during the 32-month period. The next category, "Interviewed in first wave, left sample for reasons other than death or institutionalization" includes 21,357 persons, but approximately 9,200 of this group were dropped from the sample for budget reasons. The final category includes 10,279 persons who were not members of a SIPP household during the first wave of interviews, but who subsequently became members of a sample household.

Table D-1. **Percent Distribution: Three Categories of Sample Persons**

Characteristic	Complete set of interviews obtained[1]	Interviewed in first wave, left sample for reasons other than death or institutionalization[2]	Not a member of sample household during first wave, interview obtained in second or later waves
Total	32,391 (100.0)	21,357 (100.0)	10,279 (100.0)
SEX			
Male	47.1	48.7	51.5
Female	52.9	51.3	48.5
AGE AT FIRST INTERVIEW			
Under 18 years	28.8	27.7	35.5
Under 6 years	9.9	9.3	22.9
18 to 24 years	10.7	14.2	24.6
25 to 44 years	28.5	29.5	25.4
45 to 64 years	19.5	19.3	11.2
65 years and over	12.5	9.3	3.4
75 years and over	4.8	3.2	1.4
PROGRAM PARTICIPATION: FIRST MONTH IN SAMPLE			
Persons 18 years and over	23,049 (100.0)	15,447 (100.0)	6,630 (100.0)
Participated in major assistance program	9.5	9.3	10
AFDC or general assistance	2.3	2.9	3.2
Food stamps	5.4	5.4	4.9
Medicaid	5.5	5.2	5.9
Public/subsidized.housing	3.0	3.0	2.3
SSI	2.1	1.4	1.8
Did not participate	90.5	90.7	90.0

[1] Includes 1,307 persons who died or were institutionalized during the 32-month period.
[2] Includes approximately 9,200 persons who were dropped from the panel for budget reasons.

A comparison of the first two columns shows that the characteristics of those who completed the full set of interviews are reasonably close to the characteristics of those who dropped out or were removed from the sample. The major differences in the age distribution are for young adults and the elderly. Young adults are underrepresented and the elderly are overrepresented in the group of persons who completed the full set of interviews. The data in table D-1 are, as noted, unweighted, and any potential problem caused by unrepresentative age distributions are minimized when the file is weighted to independent controls.

Time-in-Sample Bias

The use of the panel file to obtain estimates for 1984 and 1985 raises the issue of time-in-sample bias. There is ample evidence that certain measures vary according to the number of times the respondent has been visited. In CPS, for example, the measured unemployment rate is always higher for the group of households being interviewed for the first time than for the groups being interviewed for the second or later times.

Time-in-sample bias arises when a person's response to a survey question (or the interviewer's method of asking a question) is influenced by what occurred in a previous visit. The overlapping SIPP sample design will eventually provide the data that will allow for an examiniation of the presence of time-in-sample bias in SIPP estimates. That is, it is possible in SIPP to obtain estimates for a given time period from two or more separate panels and the amount of time respondents will have spent in the SIPP panel will differ for each of the panels. For example, estimates for each of the four quarters of 1985 can be obtained from both the 1984 and 1985 panels (the 1984 panel will have had more visits), and estimates for the first two quarters of 1986 can be obtained from the 1984 panel, the 1985 panel, and the 1986 panel.

Only a portion of the 1985 panel has been processed, so we have a very limited ability to examine estimates for a given time period that were produced from more than one SIPP panel. The data in table D-2 show certain estimates for the first quarter of 1985. One set of estimates was prepared using data collected in the fifth and sixth waves of the 1984 panel. The other set was prepared using data collected in the first and second waves of the 1985 panel.

The figures in table D-2 provide very weak evidence regarding the existence of time-in-sample bias for several reasons. First, most of the observed differences are smaller than the differences that could be explained by sampling sufficient to identify a pattern of bias. Second, a single observation is not sufficient to identify a pattern of bias. Third, differences may be attributable to attrition bias rather than time-in-sample bias. In spite of these qualifications, however, the observed relationships offer some reason to be cautious in interpreting the differences that have been presented earlier in this report—both the differences between CPS and SIPP estimates and the differences between the 1984 and 1985 estimates that were obtained from the SIPP. If the comparison is taken at face value, poverty estimates produced from data collected late in the panel may be subject to a downward bias.

Table D-2. **Estimates of Income and Program Participants from the 1984 and 1985 Panels of SIPP**

Income and program participants	Estimates for the first quarter of 1985	
	Based on waves 5 and 6 of the 1984 panel	Based on waves 1 and 2 of the 1985 panel
Median income of nonfarm households.	$1,811	$1,790
Percent of households with monthly income below $300..................	4.2	4.4
Percent of households with monthly income of $6,000 or more............	4.2	4.5
Percent of households with low monthly income.....................	12.9	13.6
Percent of persons who received food stamps............................	8.0	7.5
Percent of persons who were covered by Medicaid	7.5	7.0

Other Issues of Data Quality

Two major deteminants of the quality of income data collected in household surveys are the magnitude of missing responses and the accuracy of the responses that are provided. This appendix has been included to supply information concerning nonresponse rates for selected income questions, the average amounts of income reported in the survey or assigned in the imputation of missing responses, and the extent to which the survey figures underestimate numbers of income recipients and amounts of income received.

Nonresponse in this discussion refers to missing responses to specific questions or "items" on the questionnaire. Noninterviews or complete failure to obtain cooperation from any household member have not been considered in this examination of nonresponse rates. Adjustments to account for noninterviews are made by proportionally increasing the survey weights of interviewed households. Missing responses to specific questions are assigned a value in the imputation phase of the data processing operation.

Nonresponse is a very important factor in assessing the quality of survey data. Nonresponses to income questions cannot be considered random since experience has shown that persons with the highest nonresponse rates have reported characteristics such as education levels and occupations that, in general, differ from population averages. The most frequent causes of nonresponse are the inability of the respondent to answer the question because of either a 1) lack of knowledge or 2) refusal to answer. The first reason is especially important in situations of proxy response when one household member answers questions for another household member not present at the time of the interview. The practice of accepting proxy interviews from household members deemed "qualified" to answer is a standard procedure in the CPS and most other surveys conducted by the Census Bureau. During the fourth and fifth interview periods of SIPP, about 37 percent of the interviews were taken from proxy respondents.

Nonresponses are assigned values prior to producing estimates from the survey data. The procedure used to assign or impute responses for missing data for SIPP are of a type commonly referred to as a "hot deck" imputation method. This process assigns values reported in the survey by respondents to nonrespondents. The respondent from whom the value is taken is termed the "donor." Values from donors are stored in a matrix defined by demographic and economic data available for both donors and nonrespondents. Each cell of the matrix defines a unique combination of demographic and economic characteristics. For example, the imputation of an amount for monthly wage and salary income is based on eight different variables. These were 1) occupation, 2) sex, 3) age, 4) race, 5) educational attainment, 6) weeks worked, 7) usual hours worked per week, and 8) place of residence.

The second important determinant of data quality and probably the one examined most closely by users of the income data collected in household surveys is the accuracy of reported (and imputed) amounts. In general, household surveys have a tendency to underestimate the number of persons receiving income and the average amount received. These problems result for a variety of reasons including random response error, misreporting of sources of income, failure to report the receipt of income from a specified source, and failure to report the full amount received. The net effect of these kinds of problems is, for most income types, underestimation or underreporting of income amounts. The extent of underreporting is measured by comparing survey estimates with independently derived estimates, usually based on administrative data that are, generally, more reliable than the estimates derived from the survey. It should be noted that the independent estimates are subject to errors themselves. In addition, independent estimates do not reflect income attributable to the "underground" economy, some of which may be reported in the survey.